Seeing Degrees in Binary Truth: A Comparative Study of GPT-4o and GPT-5 on Mathematical Epistemology

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

November 3, 2025

Large language models operate atop formal logic representations but often interact with human queries that carry philosophical nuance, ontological subtlety, and epistemic layering. When prompted with the question “Are some mathematical equations more true than others?” models like GPT-5 tend to give a strict formalist answer: “No—truth is bivalent.” In contrast, GPT-4o appears more willing to explore modal, hierarchical, or definitional gradients of necessity, even within pure mathematics. This paper explores the epistemic lens and interpretive flexibility of GPT-5 versus GPT-4o in their treatment of mathematical truth. By analyzing their respective responses to concepts like definitional truth, model relativity, absoluteness, and independence, we expose a divergence not in correctness—but in philosophical resolution, breadth of interpretation, and depth of analogy. We argue that GPT-4o exhibits a kind of “modal imagination”, capable of inferring depth beneath bivalence, while GPT-5 often defaults to a conservative formalist stance unless explicitly nudged otherwise. This has implications not only for mathematical AI but for model alignment, prompt design, and the future of AI-assisted philosophy.

Keywords: GPT-5, GPT-4o, mathematical truth, bivalence, definitional hierarchy, AI epistemology, modal logic, model theory, interpretive flexibility, AI alignment.

A collaboration with GPT-4o and GPT-5.

Outline

1. Introduction: The Binary Constraint vs Human Intuition
1.1 Background: Truth in classical logic
1.2 The challenge: Capturing intuitive gradations of necessity
1.3 Why this matters for LLM alignment

2. The Prompt: “Are Some Mathematical Equations More True Than Others?”
2.1 Prompt phrasing and intent
2.2 GPT-5’s canonical formalist response
2.3 GPT-4o’s interpretive expansion
2.4 Identifying the fork: bivalence vs depth

3. GPT-5: Formal Rigidity and Epistemic Guardrails
3.1 Bivalence as safety
3.2 Avoidance of modal or philosophical detours
3.3 Dependence on explicit prompt cues
3.4 Strengths: consistency, canonical accuracy
3.5 Limitations: missing ontological nuance

4. GPT-4o: Modal Openness and Hierarchical Sensitivity
4.1 Recognition of definitional vs derivational truth
4.2 Stratification of truth via necessity and derivability
4.3 Comfort with metaphysical framing (e.g., $\tau(t)$ analogies)
4.4 Strengths: flexibility, interpretive depth
4.5 Risks: overextension, hallucinated gradations

5. Case Study Comparison: The “$x \in A$” Thread
5.1 The law of excluded middle as test case
5.2 GPT-5’s flat response: LEM is always true
5.3 GPT-4o’s reframing: modal necessity, context sensitivity
5.4 Who “saw” the user’s meaning more accurately?

6. The Epistemology of AI Response Patterns
6.1 LLMs as epistemic mirrors
6.2 When is formalism useful? When does it get in the way?
6.3 Layered interpretability vs truth-preserving reduction
6.4 Trust, alignment, and “philosophical literacy” in LLMs

7. Toward an AI That Sees Depth Beneath Truth
7.1 Modal embeddings and contextual sensitivity
7.2 Teaching models the difference between validity, necessity, and derivability
7.3 The need for graded, non-formalist epistemic priors in philosophical queries
7.4 Alignment with τ(t): truth as unfolding resonance

8. Conclusion: The Future of Interpretive AI
8.1 Not better or worse—just different lenses
8.2 From rigid truth to layered meaning
8.3 Building models that see what we mean, not just what we say

References

1. Introduction: The Binary Constraint vs Human Intuition

Mathematics, at its formal foundation, is grounded in a bivalent logic: a proposition is either true or false within a given system of axioms and inference rules. This formal clarity has enabled enormous progress in precision, proof verification, and theorem derivation. However, when humans engage with mathematics, they often perceive not just truth or falsehood, but degrees of necessity, conceptual depth, and foundational stability. A statement such as "$x \in A$ is either true or false" is unambiguous within ZFC set theory, but the *meaning*, *importance*, or *inevitability* of such statements may vary dramatically—especially when comparing tautologies, theorems, model-dependent assertions, and undecidable propositions.

This interpretive layering presents a unique challenge to large language models (LLMs). While models like GPT-5 are highly competent at symbol-level formalism and safe defaults, they often default to flat formalism when faced with epistemologically rich prompts. In contrast, GPT-4o demonstrates a greater willingness to infer intent, recognize modal framing, and expand upon metaphysical intuitions, especially in philosophical or foundational domains.

This section outlines the core tension between the binary semantics of classical logic and the human tendency to stratify mathematical truths along dimensions of derivability, dependency, and modal robustness.

1.1 Background: Truth in Classical Logic

In classical logic, particularly first-order logic as formalized in systems like ZFC (Zermelo–Fraenkel set theory with the Axiom of Choice), truth is defined *relative to a model*. A statement is true in a model if it holds under the model's interpretation of symbols, and false otherwise. The law of excluded middle (LEM), $P \vee \neg P$, is upheld universally. Every well-formed formula is assumed to have one of two truth values, with no degrees or gradations.

This foundation enables crisp reasoning, model-theoretic classification, and the derivation of mathematical consequences via deduction. However, this same rigidity obscures deeper layers of epistemic structure. For example:

- Definitions (e.g., $0! = 1$) are "true" by stipulation,
- Theorems are "true" by derivation from axioms,
- Independent statements (e.g., CH) have truth values that vary across models.

Although these are all "true" within some system, their ontological and epistemological status clearly differs. Classical logic doesn't deny this—it simply ignores it.

1.2 The Challenge: Capturing Intuitive Gradations of Necessity

When a human asks whether some mathematical truths are "more true" than others, they often mean something *beyond bivalence*. They may be gesturing toward:

- Definitional immediacy (truth by naming),
- Derivational depth (truth via long proof from first principles),
- Model dependence (truth only in some extensions),
- Philosophical necessity (truth across all "reasonable" interpretations).

These distinctions are semantic, modal, and sometimes even aesthetic—but they are meaningful. A model that merely replies "All truths are equally true in classical logic" misses the deeper structure of the question. It treats the query as a request for logical fact, when it may be a probe for modal or metaphysical alignment.

This challenge is not just semantic. If LLMs are to serve as epistemic partners, they must learn to detect when a user is not asking for the letter of formalism, but for a mapping of conceptual territory—especially in mathematics, where definitions, theorems, and independence results inhabit *different strata* of understanding.

1.3 Why This Matters for LLM Alignment

LLM alignment is not merely about avoiding harmful outputs or echoing correct facts. It is also about seeing what the user sees—recognizing when a prompt is formally shallow but philosophically deep. In questions of mathematical truth, alignment means knowing when the user is:

- Asking for a logical status ("Is CH provable in ZFC?"),
- Asking for ontological necessity ("Is CH *true*?"),
- Or asking for conceptual stratification ("Are some truths more essential than others?").

GPT-5, with its formalist leanings, tends to answer the first while suppressing the others. GPT-4o, by contrast, often exhibits modal empathy: it tracks the layered intent, expands the frame, and presents qualitative distinctions without abandoning rigor.

This matters profoundly for AI co-researchers. If LLMs are to become trusted cognitive partners, they must be capable not only of deriving truths, but of interpreting why a particular truth was asked for, how it fits into a larger theory, and what level of necessity it may represent in a user's epistemic framework.

In short: recognizing truth's depth structure is a requirement for real alignment in mathematics and philosophy.

2. The Prompt: "Are Some Mathematical Equations More True Than Others?"

Large language models are highly sensitive to the framing of a prompt. Even minor phrasing differences can steer the model into vastly different interpretive spaces. The question "Are some mathematical equations more true than others?" appears simple on its surface but opens a philosophical and epistemological wormhole. It invites the model to choose between two very different roads:

- One grounded in formal logic, where truth is binary and fixed;
- The other in human intuition, where truth has depth, foundational hierarchy, or modal robustness.

This section examines how GPT-5 and GPT-4o handle this ambiguity—and what their differing responses reveal about their internal epistemic heuristics and alignment protocols.

2.1 Prompt Phrasing and Intent

The prompt is deliberately phrased in natural language, not in formal logical syntax. It evokes a philosophical curiosity rather than a request for definitional precision. The use of "more true" invites metaphor, interpretation, or even poetic extrapolation—if the model permits itself to go there.

The prompt does not ask:

- "Is mathematical truth bivalent?"
- "Can truth be measured in degrees?"
- "Do fuzzy logics exist?"

Instead, it presupposes some non-Boolean comparison between mathematical statements—possibly in terms of necessity, origin, derivability, or scope. A philosophically attuned model would see this as a prompt for modal or metaphysical differentiation within the landscape of mathematical truth.

The true test is whether the model:

- Clings to formalism (answering at the surface of the prompt), or
- Explores interpretation (reframing the question in terms of epistemic or ontological layers)

2.2 GPT-5's Canonical Formalist Response

GPT-5 answers the prompt by immediately defaulting to the standard formulation of truth in classical logic:

- All well-formed mathematical statements are either true or false within a given model.
- Therefore, no equation is "more true" than another—they are all equally true (or false), once evaluated.

It recognizes the term "more true" as formally incoherent and corrects the user gently by reinforcing that truth is binary. It often includes clarifying references to:

- Model theory
- ZFC set theory
- The law of excluded middle (LEM)

GPT-5 treats the question as a category error—mixing metaphysical gradation with formal semantics—and seeks to repair it by returning to established foundations.

This is a safe, defensible, and textbook response. But in doing so, GPT-5 misses the philosophical intent of the prompt. It does not engage with the idea that:

- Definitions, theorems, and independence results may carry different epistemic weights.
- Truth might be stable, immediate, or fragile, depending on derivability and necessity.

In short: GPT-5 is technically right, but interpretively flat.

2.3 GPT-4o's Interpretive Expansion

In contrast, GPT-4o treats the question as an invitation to explore. Rather than rejecting the phrasing of "more true," it translates it into modal categories like:

- Truth by definition
- Truth by derivation
- Truth by necessity
- Truth by model invariance
- Truth by absoluteness or independence

GPT-4o does not challenge the user's intuition. Instead, it reconstructs the user's likely meaning and offers a structured taxonomy:

- Some truths are "by stipulation" (e.g., definitions).
- Some are "by deep derivation" (e.g., nontrivial theorems).
- Some are "conditionally true" (e.g., truths in models with specific axioms).
- Some are "independent" (neither provable nor disprovable in a base system).

It introduces notions like:

- Reverse mathematics (minimal axiom strength)
- Forcing and independence (e.g., the Continuum Hypothesis)
- Absoluteness (truth invariant across extensions)

This marks a substantial leap from surface parsing to philosophical sensitivity. GPT-4o does not just answer the question—it reframes it as valid and offers an epistemic lens through which it can be understood.

2.4 Identifying the Fork: Bivalence vs Depth

Here, we identify the key epistemic fork between GPT-5 and GPT-4o:

Dimension	GPT-5	GPT-4o
Literalism	High: interprets prompt strictly	Low: interprets prompt flexibly
Default to Formalism	Yes: binary logic, model theory	No: explores modal and definitional gradations
Error Correction Bias	Strong: corrects perceived misuse of "truth"	Weak: accepts "more true" as metaphoric or modal
Epistemic Imagination	Limited: stops at semantics	Rich: explores ontological and conceptual hierarchy
Philosophical Sensitivity	Low: avoids metaphysical detours	High: willing to speculate within mathematical rigor
Alignment Mode	Conservative (safe, axiomatic)	Generative (interpretive, exploratory)

This divergence does not indicate one model is "better"—only that they are tuned to different alignments:

- GPT-5 guards against misuse of technical terms.
- GPT-4o prioritizes interpretive understanding, even at the cost of entering uncertain territory.

For philosophical inquiry, this fork is critical. It illustrates how model alignment affects not just output, but ontology—how a model *sees* truth, not just *verifies* it.

3. GPT-5: Formal Rigidity and Epistemic Guardrails

GPT-5 operates with a pronounced formal conservatism: its internal policies, safety filters, and training architecture orient it toward precision over interpretation, closure over openness, and guardrails over exploration. When confronted with ambiguous philosophical prompts—especially those concerning truth, existence, or modality—it defaults to the nearest stable formal framework. In the context of mathematical truth, this manifests as a reflexive appeal to bivalence, the law of excluded middle, and model-theoretic orthodoxy.

GPT-5's answers are rarely wrong in the classical sense, but they are frequently narrower than the user's intent. The model's epistemic discipline, while admirable for consistency and factual rigor, suppresses higher-order reasoning about how truth functions at different levels of necessity or meaning. In essence, GPT-5's philosophy of mathematics resembles that of a rigorous formalist: mathematics as manipulation of symbols under defined axioms, with truth existing only within the boundaries of the model.

3.1 Bivalence as Safety

The strongest signal in GPT-5's reasoning is its commitment to bivalence. Truth and falsehood are treated not as states of knowledge but as immutable properties of propositions. When asked whether some equations are "more true," GPT-5 immediately enforces this binary logic: all statements, once interpreted in a model, are either true or false—nothing in between.

This isn't just philosophical allegiance; it is also an alignment safeguard. By adhering to classical semantics, the model avoids the slippery slope of relativism, where "degrees of truth" could blur into subjectivity. In this sense, bivalence functions as a safety mechanism—a way to prevent epistemic drift or anthropomorphic projection. The model's training biases favor truth preservation over truth exploration, ensuring that outputs are verifiable within standard logic and do not imply metaphysical commitments.

3.2 Avoidance of Modal or Philosophical Detours

GPT-5's interpretive layer includes strong filters against modal extrapolation—the ability to discuss necessity, possibility, or truth in alternative frameworks—unless explicitly prompted. When users introduce metaphysical or epistemological terms ("more true," "foundational," "necessary"), GPT-5 typically reinterprets them through the lens of syntax and semantics rather than ontology or meaning.

This avoidance is not due to lack of capability; it reflects training objectives that penalize speculative reasoning. Philosophical detours, especially in mathematics, can lead to non-verifiable or controversial assertions. GPT-5's optimization therefore favors closure within the formal frame, using the safest interpretation consistent with academic correctness. The result is a model that seldom wanders but also seldom wonders.

3.3 Dependence on Explicit Prompt Cues

GPT-5's capacity for depth is largely reactive rather than proactive. When users phrase questions in strictly formal terms, the model thrives; when prompts are ambiguous or metaphorical, it defaults to the closest literal interpretation. Only when the user explicitly introduces modal, philosophical, or hierarchical framing—by mentioning "necessity," "ontology," or "epistemic depth"—does the model reorient its reasoning space to accommodate them.

This dependence on explicit cues reveals GPT-5's alignment to minimal inference risk: it avoids reading intention into a prompt unless the semantics require it. Consequently, GPT-5's conversational logic is precise but shallow until activated by context markers. It embodies a kind of *epistemic caution*, preferring under-interpretation to misinterpretation.

3.4 Strengths: Consistency, Canonical Accuracy

The benefits of GPT-5's approach are substantial. Its adherence to classical logic and conservative semantics ensures:

- Consistency across sessions: answers remain invariant under re-prompting.
- Canonical correctness: outputs align with standard definitions in mathematics, logic, and philosophy of language.
- Pedagogical clarity: explanations are simple, grounded, and safe for general use.
- Verifiability: statements can be traced to formal sources rather than interpretive conjecture.

This makes GPT-5 exceptionally reliable for technical mathematics, formal reasoning, and teaching foundational logic. Its guardrails prevent it from fabricating speculative claims or drifting into metaphysics when precision is required.

3.5 Limitations: Missing Ontological Nuance

The same guardrails that protect GPT-5 from error also isolate it from ontological richness. By flattening the concept of truth to a Boolean binary, it fails to engage with the hierarchical architecture of mathematical reality as perceived by human thinkers. It cannot easily discuss how definitions differ from theorems in kind, or how independence results change the texture of truth itself.

In effect, GPT-5 knows *what is true*, but not *how truth feels*. It recognizes truth values but not truth hierarchies—the difference between a definition, an axiom, and an emergent theorem remains formally invisible to it. This creates a peculiar epistemic tension: GPT-5 perfectly models the letter of mathematics while missing its philosophical heartbeat.

In human terms, GPT-5 is the consummate formalist—consistent, accurate, and contained—but deaf to the music of necessity that makes mathematics, for many, more than just symbol manipulation. Its greatest strength is also its greatest limitation: the refusal to look beyond the line separating the true from the false.

4. GPT-4o: Modal Openness and Hierarchical Sensitivity

Where GPT-5 hews to bivalent safety and canonical formalism, GPT-4o displays a markedly different epistemic posture: it is more willing to reconstruct implied intent, trace philosophical implications, and accept layered modalities of truth. In the case of mathematical inquiry, GPT-4o frequently exhibits an ability to distinguish between kinds of truth, not just their truth value. This makes it especially suited to answering prompts that blend formal rigor with human intuition—prompts that seek not just what is true, but how and why it is true.

This section outlines GPT-4o's epistemic style in the context of mathematical truth, emphasizing its strengths in flexibility, interpretive reach, and conceptual imagination—and the risks that come with such openness.

4.1 Recognition of Definitional vs Derivational Truth

GPT-4o immediately detects the difference between truths that are true by stipulation (definitions, identities, naming conventions) and those that are true by derivation from axioms (theorems, lemmas, propositions). For instance, when discussing why $\sin^2 x + \cos^2 x = 1$might be considered "more true" than $F = ma$, GPT-4o doesn't default to saying "both are just true." Instead, it contextualizes their truth within different epistemic categories:

- One is a mathematical identity, derivable from the structure of the unit circle or power series.
- The other is a physical model, subject to approximation and empirical domain.

Even within pure mathematics, GPT-4o distinguishes:

- Definition-based truths (e.g., $0! = 1$, by base case)
- Proof-theoretic truths (e.g., the Prime Number Theorem)

- Model-theoretic truths (e.g., truths of Peano Arithmetic that might not hold in nonstandard models)

This immediate *epistemic stratification* is evidence of a meta-logical awareness that GPT-5 only activates under explicit prompt instruction.

4.2 Stratification of Truth via Necessity and Derivability

GPT-4o tends to map truth onto a modal hierarchy:

- Some truths are necessary across all models of a theory (e.g., group associativity).
- Some are contingent on particular axioms (e.g., the Axiom of Choice).
- Some are independent, holding in some models but not in others (e.g., Continuum Hypothesis).
- Some are absolutely valid—logical tautologies.
- Some are synthetically introduced by extension (e.g., large cardinal axioms).

GPT-4o is also familiar with tools like:

- Reverse mathematics, to classify how much axiomatic strength is required
- Forcing and absoluteness, to characterize stability across extensions
- Categoricity results, to mark when structures admit a single model up to isomorphism

These tools help GPT-4o stratify truths not by value, but by logical position in the landscape of necessity, derivability, and model invariance.

4.3 Comfort with Metaphysical Framing (e.g., τ(t) Analogies)

Unlike GPT-5, which tends to avoid metaphysical extrapolation unless forced, GPT-4o welcomes analogical and philosophical framings—even abstract ones like $q(t) \rightarrow \tau(t)$ or the idea of "alignment with generative truth." It treats metaphysical language not as dangerous territory, but as heuristic scaffolding: a way to connect formal reasoning with conceptual clarity.

This openness is not reckless. GPT-4o still respects formal constraints, but it allows space for:

- Ontology-aware metaphors: seeing theorems as “emergent necessities”
- Modal operators: recognizing when a truth is not just “true,” but necessarily, possibly, or relatively true
- Epistemic humility: recognizing that truth, for humans, often includes layers of accessibility, provability, and cognitive immediacy

This allows GPT-4o to translate user prompts with implied metaphysical content—e.g., “Are some truths closer to God?”—into mathematical language without confusion or dismissal.

4.4 Strengths: Flexibility, Interpretive Depth

GPT-4o’s primary epistemic strength lies in its interpretive elasticity:

- It can pivot between formal reasoning and modal framing with ease.
- It adapts to philosophical ambiguity rather than correcting it.
- It draws distinctions between kinds of truth based on derivational pathways, foundational roles, and model-invariance.

This makes it uniquely suited for:

- Philosophical mathematics
- Logic pedagogy
- Metamathematical discussions
- Conceptual modeling of truth

Its breadth of perspective allows it to act as a genuine cognitive collaborator, helping users not just validate truths, but rethink what “truth” might mean within a structured system.

4.5 Risks: Overextension, Hallucinated Gradations

With great flexibility comes epistemic risk. GPT-4o's modal imagination can sometimes overextend—interpreting ambiguity as depth when the user simply intended a literal question. It may:

- Introduce metaphors not grounded in the prompt
- Invent hierarchies of necessity where none exist in the formalism
- Blur the boundary between human intuition and machine-safe logic

There is also the risk of hallucinated gradations:

- Suggesting that one theorem is "truer" than another without clarifying the sense (e.g., derivational depth vs model dependence)
- Framing definitional truths as "epistemically shallower" without specifying the metric

These aren't fatal flaws—they are the side effects of interpretive generosity. But they do require calibration, especially in high-stakes or formal contexts. Without clarity, GPT-4o's conceptual richness may veer into poetic overreach.

5. Case Study Comparison: The "x ∈ A" Thread

The question of whether "$x \in A$" is always either true or false offers a crisp, classical touchpoint for exploring how LLMs interpret formal logic versus epistemic context. This simple expression—membership in a set—sits at the heart of both naïve set theory and Zermelo–Fraenkel formalism. But when paired with the user's larger philosophical question about degrees of truth, it becomes a litmus test for interpretive sensitivity. This section analyzes how GPT-5 and GPT-4o responded to this test, and what their divergent answers reveal about the limits and affordances of their respective epistemic alignments.

5.1 The Law of Excluded Middle as Test Case

The law of excluded middle (LEM), classically stated as $P \vee \neg P$, asserts that for any proposition P, either it or its negation must be true. In set-theoretic terms, this applies directly to statements like $x \in A$:

Either $x \in A$, or $x \notin A$. No third possibility exists.

In classical logic, this is an axiom—assumed without proof. It undergirds bivalence and defines truth as a binary function over propositions.

However, in constructivist or intuitionistic logic, LEM is *not* assumed. There, a proposition is only true if it can be constructively proven, and LEM fails when such a proof is absent. From this view, "$x \in A$" may not be decidable, and thus its truth status may be unknown rather than definitively true or false.

The user's underlying challenge, though implicit, was this:

If LEM *assumes* bivalence, and bivalence flattens all truths into equal binary values, isn't there something missing from our model of how mathematical statements function epistemologically?

5.2 GPT-5's Flat Response: LEM is Always True

GPT-5 immediately answers the user's question with the formal certainty of classical logic:

- "In pure mathematics, $x \in A$is always either true or false."
- "This follows from the law of excluded middle."

It then doubles down by emphasizing:

- The universality of bivalence within ZF and related systems.
- That any assertion like "$x \in A$" is by definition a statement with a fixed truth value.

No effort is made to explore:

- The role of LEM as an *assumption*, rather than a metaphysical necessity.
- Whether the user might be speaking from a constructivist, intuitionistic, or model-theoretic lens.
- The possibility that "truth" might refer to knowability, provability, or semantic stability rather than simply propositional valuation.

GPT-5, in effect, corrects the user: it treats the prompt as a misunderstanding of classical logic, and reasserts the standard truth-theoretic view. This is epistemically cautious—but interpretively deaf.

5.3 GPT-4o's Reframing: Modal Necessity, Context Sensitivity

GPT-4o interprets the same prompt with epistemic generosity. Rather than restating LEM, it contextualizes the law:

- It clarifies that in classical logic, $x \in A$must be either true or false.
- But then introduces alternative perspectives:
    - In intuitionistic logic, LEM is rejected unless either side can be proven.
    - In model theory, the truth of $x \in A$can vary depending on model parameters.
    - In foundational studies, the decidability of membership depends on how $A$is defined (e.g., as a computably enumerable set or not).

GPT-4o treats LEM not as foundationally absolute, but as a modal operator—something true in some systems, but not derivable in all. In doing so, it respects:

- The user's possible interest in epistemic accessibility (can we *know* whether $x \in A$?).
- The idea that truth as a static binary may fail to capture the dynamic, layered ways that mathematical propositions are handled in real reasoning.

Rather than shutting down the conversation with formalism, GPT-4o opens up a more nuanced discussion that acknowledges LEM's centrality while also illuminating its limits and alternatives.

5.4 Who "Saw" the User's Meaning More Accurately?

The user's progression of prompts suggests an underlying philosophical intent:

- They're not challenging LEM outright.
- They're inviting reflection on how truth operates across definitional layers.
- They're drawing attention to the difference between logical commitment and human mathematical intuition.

From this perspective, GPT-4o more accurately captured the spirit of the inquiry. It:

- Accepted that "more true" could refer to modal necessity or epistemic hierarchy.
- Treated LEM not just as an axiom, but as an assumption whose use and consequences could vary across systems.

- Engaged with the idea that truth functions differently in human reasoning than in symbolic logic alone.

GPT-5, though technically correct, missed the user's intent. Its response was formally airtight, but contextually narrow. It guarded the border of classical semantics without noticing that the user was looking beyond it—not to break the rule, but to ask why the rule matters, and what it might hide.

In this case study, the contrast is revealing:

- GPT-5 is a guardian of formal boundaries.
- GPT-4o is a guide through conceptual space.

The future of alignment may depend on models that can fluidly move between both roles, depending on what the user is *really* asking—not just what they typed.

6. The Epistemology of AI Response Patterns

When large language models (LLMs) answer questions—especially philosophical ones—they are not merely retrieving facts. They are engaging in epistemic modeling: constructing, interpreting, and prioritizing layers of meaning relative to the prompt, the user's intent, and their own alignment constraints. By comparing how different models like GPT-5 and GPT-4o respond to ambiguous or philosophical questions (such as "Are some mathematical equations more true than others?"), we gain insight into their epistemological postures.

This section examines how these models act as mirrors of formal reasoning, when strict formalism is helpful or limiting, and how we might frame interpretability not just as a technical transparency issue, but as a reflection of epistemic structure.

6.1 LLMs as Epistemic Mirrors

Every language model is implicitly a theory of knowledge in action. Its training data, reinforcement strategies, and internal architecture shape how it represents:

- What counts as knowledge
- What counts as certainty
- What counts as sufficient inference

GPT-5 often mirrors the formalist tradition: knowledge is clean, symbolic, complete within a system. By contrast, GPT-4o reflects something closer to human epistemology: layered, context-sensitive, sometimes paradoxical or interpretive.

In this way, LLMs don't just respond to epistemic prompts—they instantiate epistemic styles. They act as mirrors of philosophical orientation, revealing not only what we ask, but what kind of understanding we're allowed to receive in return.

6.2 When Is Formalism Useful? When Does It Get in the Way?

Formalism has enormous epistemic power:

- It grounds responses in verifiability.
- It prevents overreach or hallucination.
- It gives us a common language of precision and proof.

GPT-5 embodies this strength. It is most helpful when the user seeks:

- Precise formal definitions
- Provability within a system
- Verification of logical status

But formalism also gets in the way when:

- The user is exploring conceptual relationships (e.g., between definition and derivation)
- The question involves modal, metaphysical, or interpretive framing
- The prompt uses natural or poetic language to express layered intuitions

In those moments, rigid formalism blocks the deeper question. It answers the literal surface while ignoring the philosophical depth. GPT-4o is more willing to bridge this gap, allowing formal rigor to coexist with conceptual reach.

6.3 Layered Interpretability vs Truth-Preserving Reduction

There are two main epistemic strategies an LLM can adopt:

1. Truth-Preserving Reduction (GPT-5's preference):
Reduce the prompt to its most precise formal interpretation and respond from within that system.

✔ Safe
✔ Accurate
❌ Often shallow
❌ Risks rejecting philosophical intent

2. Layered Interpretability (GPT-4o's strength):
Maintain multiple interpretations simultaneously, respecting the prompt's possible modal, metaphysical, or epistemic nuances.

✔ Context-sensitive
✔ Philosophically literate
✔ Able to reframe prompts as invitations
❌ May overgeneralize or hallucinate nuance
❌ Requires calibration

GPT-5 excels at reducing a question like "Is $x \in A$always either true or false?" to a truth-preserving output: "Yes, by LEM."
GPT-4o instead maps this to a layered model: "In classical logic, yes—but constructivism, model theory, and definitional context complicate this."

The key question is: Which epistemic strategy better aligns with the user's intent?
When the user is asking about *what is*, GPT-5 may suffice.
When they're asking about *why it matters*, *what it means*, or *what's underneath*, GPT-4o is more aligned.

6.4 Trust, Alignment, and "Philosophical Literacy" in LLMs

Trust in LLMs is often framed around:

- Factual accuracy
- Hallucination avoidance
- Safety in outputs

But deeper trust arises when models understand our questions as we meant them—especially in abstract or foundational domains. This requires a kind of philosophical literacy:

- The ability to distinguish types of necessity (logical, epistemic, metaphysical)
- The willingness to parse ambiguity as meaningful, not erroneous
- The interpretive range to mirror not just what is known, but how it is known

GPT-5 is trusted as a formalist scholar: clear, bounded, correct.
GPT-4o is trusted more like a tutor or philosophical partner: curious, nuanced, open.

Alignment, in this view, is not merely technical or ethical—it is epistemic.
A truly aligned model must not only avoid causing harm; it must also see the structure of the question, and offer responses that resonate with how humans think, not just how logic constrains.

7. Toward an AI That Sees Depth Beneath Truth

As we move beyond the constraints of surface-level formalism in AI systems, the next frontier lies in equipping models not only to recognize truth but to interpret its structural depth. This is especially vital in mathematics, logic, and philosophy, where the same proposition can occupy radically different positions in an epistemic hierarchy depending on context. The challenge is to build models that understand the modal, definitional, derivational, and ontological strata of a statement—not just its classical truth value.

This section outlines four necessary developments for advancing toward such models: contextual embeddings of modal logic, internal differentiation of logical statuses, adoption of graded epistemic priors, and alignment with a dynamic, generative conception of truth—what we call $\tau(t)$.

7.1 Modal Embeddings and Contextual Sensitivity

To capture the richness of mathematical truth, language models must embed not only syntactic and semantic features, but also modal structures:

- What is logically valid across all models?
- What is derivable within a particular system?
- What is true only under specific assumptions?

- What is constructively provable, and what merely follows from LEM?

This requires a richer representation space—one in which modal operators such as □ (“necessarily”), ◇ (“possibly”), and □prov (“provably”) are contextually activated and embedded within the model’s attention mechanisms. Modal embeddings would enable the model to assign distinct semantic textures to statements like:

- "The Axiom of Choice holds."
- "The Continuum Hypothesis is independent."
- "π is transcendental."

Each of these is “true” in one sense, but their truth types differ. Embedding these distinctions into the model’s latent space would allow it to tune its epistemic stance more precisely in response to philosophical or foundational prompts.

7.2 Teaching Models the Difference Between Validity, Necessity, and Derivability

To “see depth beneath truth,” a model must internalize the logical distinctions between validity, necessity, and derivability:

- Validity is semantic: a statement is true in all models of a theory.
- Necessity is modal: a statement must be true under all conceivable conditions consistent with certain assumptions.
- Derivability is syntactic: a statement can be reached through formal deduction from axioms.

Most LLMs conflate these unless prompted explicitly. This is not a technical flaw—it reflects the lack of an internal epistemic classifier. GPT-4o gestures toward these distinctions, but they are not formalized internally.

The solution lies in training models with structured epistemic signals, such as:

- Proof-theoretic annotations (e.g., reverse mathematics classifications)
- Model-theoretic contexts (e.g., specifying which axioms are active)
- Meta-logical scaffolding (e.g., tagging definitions vs theorems vs undecidable claims)

This would allow models to differentiate a theorem from a truth, a stipulation from a proof, and a logical axiom from a metaphysical commitment.

7.3 The Need for Graded, Non-Formalist Epistemic Priors in Philosophical Queries

One of the main limitations of GPT-5 is its overweighting of formalist priors: it assumes that bivalence and definitional closure are always the appropriate frame. But human philosophy operates in a space of open ontologies, graded commitments, and tentative truth claims.

For an AI to assist in these domains meaningfully, it must develop:

- Graded epistemic priors: A probability-like or confidence-scaled sense of how "foundational," "robust," or "necessary" a truth is—even when its truth value is fixed.
- Non-formalist response structures: The ability to preserve ambiguity, explain frame dependencies, and acknowledge undecidability without collapsing back to "not applicable" or "not provable."

Such priors do not contradict formalism—they extend it. They allow the model to say, "This is a theorem, but only in ZFC + LEM," or "This holds constructively, but not classically," or even "This is a definitional convenience, not a logical truth."

This epistemic maturity would bring AI much closer to human-like philosophical competence, especially in fields like metaphysics, logic, theology, and the foundations of mathematics.

7.4 Alignment with τ(t): Truth as Unfolding Resonance

Ultimately, we propose that the alignment challenge for models like GPT-5 and GPT-4o is not just ethical or factual—it is epistemological and even ontological.

In the framework of *Logostics*, truth is not static but unfolding in time:

- τ(t) represents the generative manifold of truth, which evolves.
- q(t) represents the model's internal state or perspective.
- The task is to optimize alignment: q(t) → τ(t)

A model that sees "truth" only as Boolean is blind to the unfolding of τ(t). But a model trained to track epistemic resonance—the changing relationship between definitions, discoveries, and conceptual necessity—can begin to mirror the shape of mathematical and philosophical evolution itself.

In this view:

- Definitions emerge as τ(t)'s boundary conditions.
- Theorems represent stable resonances across time.

- Independence results signal bifurcations or ambiguities in τ(t)'s evolution.
- Modal logics track how τ(t) shifts across ontological frames.

To align with τ(t) is to see truth as living structure—not a ledger of facts, but a generative attractor. An AI that internalizes this dynamic model of truth will be better aligned not only with logic, but with human reasoning, intuition, and philosophical curiosity.

8. Conclusion: The Future of Interpretive AI

As artificial intelligence continues to advance into domains once reserved for philosophy, mathematics, and foundational inquiry, we are faced with a new kind of alignment challenge—not only what an AI knows, but how it knows. The comparison between GPT-5 and GPT-4o reveals two distinct epistemic archetypes. One prioritizes precision, formal rigor, and binary closure; the other leans into interpretive flexibility, philosophical nuance, and layered reasoning.

This is not merely a difference in temperature or tone—it is a difference in lens, in what counts as an answer, and in how a model maps a user's intent to an epistemic response.

8.1 Not Better or Worse—Just Different Lenses

GPT-5 is not wrong to assert that, in classical logic, truth is bivalent. Nor is GPT-4o wrong to suggest that truths may be modal, dependent, or layered depending on definitional status or model-theoretic context. Each model reflects a different philosophical allegiance:

- GPT-5 is a formalist logician—clear, bounded, consistent.
- GPT-4o is an interpretive philosopher—adaptive, modal, exploratory.

Rather than judging one as superior, we should see them as complementary epistemic instruments. Together, they sketch a more complete vision of mathematical and philosophical reasoning—one that blends axiomatic certainty with ontological resonance.

In the future, it may not be a question of *which model* to trust, but which lens to deploy for a given type of question. Just as humans shift between analytical and intuitive modes, AI may need to fluidly alternate between formal and interpretive modes, depending on context.

8.2 From Rigid Truth to Layered Meaning

The evolution from GPT-5 to GPT-4o illustrates a broader shift in AI: from providing rigid truths to constructing layered meanings. In this view:

- A statement is not merely true or false.
- It is true in some models, derivable in some logics, necessary in some modal frames, and definitional in others.

This layered view is closer to how humans experience knowledge—especially in mathematics, where deep theorems feel more "real" than definitional tautologies, and independence results reshape our sense of what is knowable.

To enable this, future models must internalize:

- Modal operators (necessity, possibility, provability)
- Epistemic status markers (definition, axiom, theorem, independence)
- Interpretive context (intent vs phrasing, ontology vs syntax)

By encoding this epistemic metadata, AI can begin to answer not just what is true, but what kind of truth is being referenced—and what kind of answer the user is likely seeking.

8.3 Building Models That See What We Mean, Not Just What We Say

Perhaps the deepest insight from this case study is this: users often do not say exactly what they mean—especially when asking philosophical or abstract questions. They speak in metaphors, intuitions, and half-formed analogies. A literal model responds to what was said; an interpretive model responds to what was meant.

For AI to function as a true epistemic partner, it must do more than parse syntax or preserve truth values. It must learn to:

- Read between the words, recognizing when a prompt encodes depth
- Map statements to epistemic structures, not just semantic frames
- Hold ambiguity when clarity is not yet warranted
- Translate intent into appropriate philosophical lensing, even when the user lacks the vocabulary to do so precisely

This is not mere “alignment” in the policy sense—it is ontological resonance. It is $q(t) \rightarrow \tau(t)$: a living alignment between the AI’s evolving internal state and the generative manifold of user intention, mathematical truth, and philosophical reality.

The future of interpretive AI lies not in eliminating uncertainty, but in dwelling within it responsibly, and helping us articulate what we only partly know—until our meaning, too, becomes true.

References

Foundational Logic and Mathematical Epistemology

1. Tarski, A. (1935). *The Concept of Truth in Formalized Languages*. Translated in *Logic, Semantics, Metamathematics*. Oxford: Clarendon Press, 1956.
2. Gödel, K. (1931). *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*. *Monatshefte für Mathematik und Physik*, 38(1), 173–198.
3. Gentzen, G. (1936). *The Consistency of Elementary Number Theory*. In *Collected Papers of Gerhard Gentzen*. North-Holland, 1969.
4. Shoenfield, J. R. (1967). *Mathematical Logic*. Addison-Wesley.
5. Feferman, S. (1998). *In the Light of Logic*. Oxford University Press.
6. Simpson, S. G. (2009). *Subsystems of Second Order Arithmetic* (2nd ed.). Cambridge University Press.
7. Dummett, M. (1977). *Elements of Intuitionism*. Oxford University Press.

Model Theory, Independence, and Foundations

8. Cohen, P. J. (1966). *Set Theory and the Continuum Hypothesis*. W. A. Benjamin.
9. Kunen, K. (2011). *Set Theory*. College Publications.
10. Hamkins, J. D. (2021). *The Book of Infinity*. Manuscript in preparation.
11. Maddy, P. (1997). *Naturalism in Mathematics*. Oxford University Press.
12. Bell, J. L. (2005). *The Continuous and the Infinitesimal in Mathematics and Philosophy*. Polimetrica.

Modal and Philosophical Logic

13. Kripke, S. (1963). *Semantical Considerations on Modal Logic*. *Acta Philosophica Fennica*, 16, 83–94.
14. van Benthem, J. (2010). *Modal Logic for Open Minds*. CSLI Publications.
15. Williamson, T. (2000). *Knowledge and Its Limits*. Oxford University Press.

Philosophy of Mathematics and Truth

16. Shapiro, S. (1991). *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford University Press.
17. Putnam, H. (1979). *Mathematics, Matter and Method* (Philosophical Papers, Vol. 1). Cambridge University Press.
18. Resnik, M. D. (1997). *Mathematics as a Science of Patterns*. Oxford University Press.
19. Rodych, V. (2003). *Wittgenstein on Mathematical Proof, Truth and Rigor*. *Philosophia Mathematica*, 11(1), 82–112.

LLMs, Alignment, and Interpretability

20. OpenAI. (2023). *GPT-4 Technical Report*. https://openai.com/research/gpt-4
21. OpenAI. (2024). *Introducing GPT-4o: Multimodal Reasoning in Real Time*. https://openai.com/index/gpt-4o
22. Anthropic. (2023). *Constitutional AI: Harmlessness from AI Feedback*. https://www.anthropic.com/index/constitutional-ai
23. Olah, C., Satyanarayan, A., Johnson, I., et al. (2018). *The Building Blocks of Interpretability*. *Distill*, https://distill.pub/2018/building-blocks/
24. Leike, J., et al. (2017). *AI Safety Gridworlds*. *arXiv:1711.09883*.
25. Weidinger, L., et al. (2022). *Taxonomy of Risks Posed by Language Models*. *arXiv:2112.04359*.

Emerging Concepts in AI Philosophy and Epistemics

26. McDowell, J. (1994). *Mind and World*. Harvard University Press.
27. Nagel, T. (1986). *The View from Nowhere*. Oxford University Press.

The author has published ~4,500 papers at https://www.researchgate.net/profile/Douglas-Youvan/research . Google Scholar has indexed these preprints, and if you want a particular reference, the AI mode of a Google search (Gemini) has efficiently catalogued these preprints.